Automatic calibration of an ensemble for uncertainty estimation and probabilistic forecast: Application to air quality

Damien Garaud, Vivien Mallet

Automatic calibration of an ensemble for uncertainty estimation and probabilistic forecast: Application to air quality

D. Garaud[1] and V. Mallet[2]

Received 8 February 2011; revised 25 May 2011; accepted 21 July 2011; published 13 October 2011.

[1] This paper addresses the problem of calibrating an ensemble for uncertainty estimation. The calibration method involves (1) a large, automatically generated ensemble, (2) an ensemble score such as the variance of a rank histogram, and (3) the selection based on a combinatorial algorithm of a sub-ensemble that minimizes the ensemble score. The ensemble scores are the Brier score (for probabilistic forecasts), or derived from the rank histogram or the reliability diagram. These scores allow us to measure the quality of an uncertainty estimation, and the reliability and the resolution of an ensemble. The ensemble is generated on the Polyphemus modeling platform so that the uncertainties in the models' formulation and their input data can be taken into account. A 101-member ensemble of ground-ozone simulations is generated with full chemistry-transport models run across Europe during the year 2001. This ensemble is evaluated with the aforementioned scores. Several ensemble calibrations are carried out with the different ensemble scores. The calibration makes it possible to build 20- to 30-member ensembles which greatly improves the ensemble scores. The calibrations essentially improve the reliability, while the resolution remains unchanged. The spatial validity of the uncertainty maps is ensured by cross validation. The impact of the number of observations and observation errors is also addressed. Finally, the calibrated ensembles are able to produce accurate probabilistic forecasts and to forecast the uncertainties, even though these uncertainties are found to be strongly time-dependent.

Citation: Garaud, D., and V. Mallet (2011), Automatic calibration of an ensemble for uncertainty estimation and probabilistic forecast: Application to air quality, *J. Geophys. Res.*, *116*, D19304, doi:10.1029/2011JD015780.

1. Introduction

[2] Air quality simulation involves complex numerical models that rely on large amounts of data from different sources. Most of the input data is provided with high uncertainties in their time evolution, spatial distribution and even average values. Chemistry-transport models are themselves subject to uncertainties in both their physical formulation and their numerical formulation. The multi-scale nature of the problem leads to the introduction of subgrid parameterizations that are an important source of errors. The dimensionality of the numerical system, involving up to hundreds of pollutants in a three-dimensional mesh, is much higher than the number of observations, which also leads to high uncertainties in non-observed variables.

[3] In order to quantify the uncertainties, classical approaches rely on Monte Carlo simulations. The input fields and parameters of the chemistry-transport model are viewed as random vectors or random variables. These are sampled according to their assumed probability distribution, and a model run is carried out with each element of the sample. The set of model outputs constitutes a sample of the probability distribution function of the output concentrations. Typically, the empirical standard deviation of the output concentrations measures the simulations uncertainties. This approach has been applied for air quality simulations [*Hanna et al.*, 1998, 2001; *Beekmann and Derognat*, 2003].

[4] Another approach is the use of models which differ by their numerical formulation or physical formulation. The models can originate from different research groups [e.g., *van Loon et al.*, 2007; *Delle Monache and Stull*, 2003; *McKeen et al.*, 2005; *Vautard et al.*, 2009] or from the same modular platform [*Mallet and Sportisse*, 2006]. In addition to this multimodel strategy, the input data can also be perturbed so that all uncertain sources are taken into account. It is also possible to choose between different emission scenarios and meteorological forecasts as *Delle Monache et al.* [2006a, 2006b] did. *Pinder et al.* [2009] split the uncertainty into a structural uncertainty due to the weaknesses in the physical formulation and a parametric uncertainty due to the errors in the input data. *Garaud and Mallet* [2010] built the ensemble with several models randomly generated within the same platform and with perturbed input data.

[1]CEREA, Joint Research Laboratory ENPC ParisTech–EDF R&D, Université Paris-Est, Marne-La-Vallée, France.
[2]Paris-Rocquencourt Research Center, INRIA, Le Chesnay, France.

0148-0227/11/2011JD015780

[5] Whatever the strategy for the generation of an ensemble, several assumptions are made by the modelers. One needs to associate probability density functions to every input field or parameter to be perturbed. Under the usual assumption that the distribution of a field or parameter is either normal—or log-normal, one has to estimate a median and a standard deviation. For a field, providing a standard deviation is complex as it should take into account spatial correlations, and possibly time correlations. As for multimodel ensembles, one has little control over the composition of the models when they are provided by different teams. When the models are derived within the same platform, the key points are the amount of choice in the generation of an individual model, and the probability associated to each choice. Once all the assumptions and choices have been made, it is technically possible to generate an ensemble. However, it is quite difficult to determine the proper medians and standard deviations of the perturbed fields, and to design a multimodel ensemble that properly takes into account all formulation uncertainties.

[6] In order to evaluate the quality of an ensemble, several a posteriori scores compare the ensemble simulations with observations. These scores, such as rank histograms, reliability diagrams or Brier scores, assess the reliability, the resolution or the sharpness of an ensemble. For instance, a reliable ensemble gives a well estimated probability for a given event in comparison to the frequency of occurrence of this event, whereas the resolution describes the capacity of an ensemble to give different probabilities for a given event.

[7] Improving the quality of an ensemble should lead to improved scores, e.g., to a flat rank diagram or low Brier score. One strategy could be tuning the perturbations of the input fields or optimizing the design of the multimodel ensemble (that is, choosing or developing physical parameterizations or numerical schemes, and better weighting each design option), so as to minimize or maximize some score. This is a complex and computationally expensive task that would require the generation of many ensembles.

[8] In this paper, we adopt a strategy based on a single, but large, ensemble. Out of a large ensemble, a combinatorial optimization algorithm extracts a sub-ensemble that minimizes (or maximizes) a given score such as the variance of a rank diagram. This process is referred to as (a posteriori) calibration of the ensemble. Section 2 describes it in detail. It is applied in Section 3 to a 101-member ensemble of ground-ozone simulations with full chemistry-transport models run across Europe during the year 2001. The scores of the full ensemble and the optimized sub-ensemble (i.e., the calibrated ensemble) are studied, based on observations at ground stations. In Section 4, the uncertainty estimation given by the calibrated ensemble is analyzed. In Section 5, probabilistic forecasts for threshold exceedance are studied.

2. Calibration Method

[9] *Hamill and Colucci* [1997] use rank histograms to calibrate precipitation probabilistic forecasts. When the ensemble is not reliable enough, the probabilistic forecasts cannot be derived directly from the ensemble relative frequencies. Assuming the shape of the rank histogram remains the same in the forecast period, the authors propose to rely on the past rank distribution to compute the probabilistic forecasts. *Hopson and Webster* [2010] calibrate an ensemble prediction to improve floods forecasting. An empirical cumulative distribution function is provided by ensemble predictions of precipitation. Then, it is calibrated with observations, using a quantile-to-quantile mapping technique.

[10] In this paper, by "ensemble calibration" we mean extracting a sub-ensemble from a large ensemble so that a certain criterion is satisfied. A preliminary step is therefore to generate a large ensemble, composed of simulations that are sufficiently different from each other to provide substantial information. A criterion is defined to assess the quality of an ensemble, and a corresponding score measures how well the criterion is satisfied. An automatic selection of a sub-ensemble is finally carried out to minimize the score. The criterion usually assesses the uncertainty representation of an ensemble, based on the additional information brought by the observations. This section details the method employed to generate a large ensemble and to carry out an automatic calibration.

2.1. Generation of a Large Ensemble

[11] The method employed for the automatic generation of a large ensemble is described by *Garaud and Mallet* [2010]. A wide range of options should be available for the design of a single model: several physical parameterizations, several numerical discretizations, different sources for the input data and random perturbations in the input fields. In the paper referred to, thirty alternatives are available for the generation of a single model. Each member of the ensemble is defined after the random selection of one option per alternative.

[12] In this paper, we rely on the same ensemble as *Garaud and Mallet* [2010]. It includes 101 members run throughout the year 2001 over Europe. This ensemble will be used and calibrated in Section 3.

2.2. Automatic Selection

[13] Suppose a base ensemble with N members. There are $\sum_{k=1}^{N}\binom{N}{k}$ possible sub-ensembles. If $N = 100$, there are over 10^{30} sub-ensembles. It is obviously impossible to consider all possible combinations in order to select the best combination with respect to the given criterion. Consequently a combinatorial optimization algorithm is required to minimize the score associated with the criterion.

[14] Let $\mathcal{E}$ be the full ensemble and $\mathcal{S}$ be a sub-ensemble of $\mathcal{E}$. $\mathcal{S} \subseteq \mathcal{E}$ is supposed to be non-empty. Let $J(\mathcal{S})$ be the score of $\mathcal{S}$. The following sections describe different scores and algorithms which may be used in the ensemble calibration.

2.2.1. Criterion and Score

[15] The main reasons for generating an ensemble are to improve forecasts with the so-called ensemble forecasts, and to estimate the uncertainty in the model's output. In this paper, we focus on the second objective. The criterion typically measures the quality of an uncertainty estimation or of the prediction of exceeding a threshold. It can be based on two desirable features of an ensemble:

[16] 1. Reliability: an ensemble has high reliability when its probabilistic forecasts for a given event match, on average, the observed frequency of this event.

[17] 2. Resolution: the capacity of the prediction system to distinguish the outcomes for a given event.

2.2.1.1. Rank Histogram

[18] A rank histogram measures the reliability of an ensemble. Let $\{x_1,\ldots x_j,\ldots, x_N\}$ be the output of a N-member ensemble at a given time, sorted in increasing order. This ensemble is considered as a sample of a random variable X with some probability distribution, which means that all x_j are supposed to follow the same probability distribution. Let Y be a random variable representing the true state. At a given point, if Y has the same probability distribution as X, then $\mathrm{E}_X[P_Y(y \le x_j)] = \frac{j}{N+1}$, where $\mathrm{E}_X[.]$ denotes the expectation related to X, P_Y the probability associated with Y and y a realization of the true state, i.e., an exact measured ozone concentration for instance. The rank histogram, developed by *Anderson* [1996], *Talagrand et al.* [1999], and *Hamill and Colucci* [1997], is computed by counting the rank of the true state to an actual sorted ensemble of forecasts. A perfect diagram is flat, whereas a U-shaped rank histogram means a lack of variability in the ensemble.

[19] Let r_j be the number of observations of rank j. An observation of rank j is an observation which is higher than the concentrations of exactly j members of the ensemble. Suppose we have M observations. The expectation of r_j is $\bar{r} = \mathrm{E}[\sum_{m=1}^{M} P_Y(x_j < y_m \le x_{j+1})] = \frac{M}{N+1}$. The score related to the rank histogram flatness is based on the squared error

$$\mathcal{S} = \sum_{j=0}^{N} \left(r_j - \bar{r}\right)^2. \tag{1}$$

[20] The score $\mathcal{S}$ gets lower as the histogram gets flatter, since $\bar{r}$ corresponds to the height of a flat histogram. Obviously, this measure depends on the number of members. It can be normalized by $\mathcal{S}_0 = \mathrm{E}[\mathcal{S}] = \frac{NM}{N+1}$ because $\mathrm{E}[(r_j - \bar{r})^2] = \frac{NM}{(N+1)^2}$. Finally the following score is used to measure the flatness of the rank histogram:

$$\delta = \frac{N+1}{NM} \sum_{j=0}^{N} \left(r_j - \bar{r}\right)^2, \tag{2}$$

which should ideally be close to 1.

2.2.1.2. Reliability Diagram

[21] Instead of simply predicting whether an event will occur or not, an ensemble can provide a probabilistic forecast. This is especially useful for the prediction of a threshold exceedance. A basic probabilistic forecast may be given by the number of models which exceed the threshold over the total number of models [*Anderson*, 1996]. In order to construct a reliability diagram, the range of forecast probabilities, [0, 1], is divided into $K + 1$ bins $[p_0, p_1],\ldots,$ $[p_k, p_{k+1}], \ldots, [p_{K-1}, p_K]$ where $p_0 = 0$, $p_K = 1$, and the sequence $(p_k)_k$ is increasing. Let O_k be the (observed) relative occurrence frequency of the event when the ensemble predicts in $[p_k, p_{k+1}]$. A reliable ensemble should give $O_k \in [p_k, p_{k+1}]$. The reliability diagram [*Wilks*, 2005] plots O_k against p_k or $\frac{1}{2}(p_k + p_{k+1})$. A perfect reliability diagram should follow the diagonal.

2.2.1.3. Brier Score

[22] The Brier score measures the mean squared probability error for a specific event [*Brier*, 1950; *Wilks*, 2005]. Let M be the total number of observations. Let p_i be the forecast probability and o_i be the observed probability at a date i. The observed probability o_i is equal to 1 if the event occurred, and 0 otherwise. The Brier score is given by:

$$\mathcal{B} = \frac{1}{M} \sum_{i=1}^{M} (p_i - o_i)^2. \tag{3}$$

[23] A Brier score for an ensemble can be compared with the Brier score of the climatological forecast. The climatological forecast is given by a single occurrence frequency o_c, observed in the past. If o_i follows the Bernoulli distribution and is equal to 1 with the frequency o_c and to 0 with the frequency $1 - o_c$, the expectation of the Brier score $\mathcal{B}_{cl}$ of the climatological forecast is given by

$$\mathcal{B}_{cl} = \frac{1}{M} \sum_{i=1}^{M} \left[o_c(o_c - 1)^2 + (1 - o_c)o_c^2\right] = o_c(1 - o_c). \tag{4}$$

The so-called Brier skill score is defined by

$$\mathcal{B}_s = 1 - \frac{\mathcal{B}}{o_c(1 - o_c)}. \tag{5}$$

It ranges between $[-1, 1]$ and is greater than 0 when the ensemble prediction gives a better forecast than the climatological forecast.

2.2.1.4. Discrete Ranked Probability Score

[24] Suppose a set of L events, and let p_{li} be the forecast probability for the l-th event at the date i. The total number of observations M is the same for each event. The discrete ranked probability score (DRPS), which is a variant of RPS (ranked probability score) [*Epstein*, 1969; *Murphy*, 1971], is given by:

$$\mathrm{DRPS} = \frac{1}{LM} \sum_{i=1}^{M} \sum_{l=1}^{L} (p_{il} - o_{il})^2$$

$$\mathrm{DRPS} = \frac{1}{L} \sum_{l=1}^{L} \mathcal{B}(\mathcal{E}_l). \tag{6}$$

[25] This score is a generalization of the Brier score from a single event to a set of events.

[26] While the rank histogram and the reliability diagram measure the reliability of a prediction system, the Brier score, and thus the DRPS, can measure the reliability and the resolution of an ensemble as shown in [*Murphy*, 1973]. The latter scores can be broken down into three terms: reliability, resolution and uncertainty. For instance, the Brier score is an estimation of $\mathrm{E}[(p - o)^2]$. Let p_0 be the specific probability for a given event $\mathcal{E}$ and O_0 be the occurrence frequency of $\mathcal{E}$ when p_0 is provided. The occurrence of $\mathcal{E}$ denoted o follows Bernoulli's distribution. Thus, o takes value 1 with frequency O_0 and takes value 0 with frequency $1 - O_0$. The expected value of $(p_0 - o)^2$ is

$$\begin{aligned} \mathrm{E}\left[(p_0 - o)^2\right] &= (p_0 - 1)^2 O_0 + p_0^2(1 - O_0) \\ &= (p_0 - O_0)^2 + O_0(1 - O_0)\ . \end{aligned} \tag{7}$$

[27] Then, we compute (7) for many probabilities. In our case, the prediction system provides discrete probabilities for a given event. Suppose the system provides $K + 1$ different probabilities denoted p_k, ranging in [0, 1]. Let n_k be the number times p_k is computed with the ensemble. Thus, the frequency distribution of p_k is given by $\frac{n_k}{M}$ with M the total number of considered dates, i.e., the total number of observations. We have $\frac{1}{M}\sum_{k=0}^{K} n_k = 1$. Let O_k be the observed occurrence frequency of the event when the ensemble predicts p_k. The climatological occurrence frequency is $o_c = \frac{1}{M}\sum_{k=0}^{K} n_k O_k$.

$$\begin{aligned}\mathcal{B} &= \mathrm{E}\left[(p-o)^2\right] \\ &= \frac{1}{M}\sum_{k=0}^{K} n_k (p_k - O_k)^2 + \sum_{k=0}^{K} n_k O_k (1 - O_k) \\ &= \underbrace{\frac{1}{M}\sum_{k=0}^{K} n_k (p_k - O_k)^2}_{\text{reliability}} - \underbrace{\sum_{k=0}^{K} n_k (O_k - o_c)^2}_{\text{resolution}} + \underbrace{o_c(1-o_c)}_{\text{uncertainty}}.\end{aligned} \tag{8}$$

[28] The first term is a reliability term since it compares the probability provided by the forecast system with the occurrence frequency of the event. The second term is called "resolution" and is equivalent to the variance of O_k. The third one is the "uncertainty" term which corresponds to the score of the climatological forecast. It is constant for a specific event and is maximum when the climatological forecast is equal to 0.5. This means that the climatological forecast has the worst Brier score when it provides the most uncertain occurrence probability, i.e., 0.5. The same decomposition can be carried out for the Brier skill score and the DRPS (9).

$$\begin{aligned}DRPS = &\frac{1}{LM}\sum_{l=1}^{L}\sum_{k=0}^{K} n_k^l (p_{lk} - O_{lk})^2 \\ &- \frac{1}{LM}\sum_{l=1}^{L}\sum_{k=0}^{K} n_k^l (O_{lk} - o_{lc})^2 \\ &+ \frac{1}{L}\sum_{l=1}^{L} o_{lc}(1 - o_{lc})\end{aligned} \tag{9}$$

[29] The choice of a criterion, i.e., an ensemble score, is the first step of the ensemble calibration. The second step is the choice of a combinatorial optimization algorithm.

2.2.2. Combinatorial Optimization Algorithm

[30] Two combinatorial optimizations are employed in order to minimize the scores previously introduced: a genetic algorithm and simulated annealing.

2.2.2.1. Genetic Algorithm

[31] The genetic algorithm, described by *Fraser and Burnell* [1970] and *Crosby* [1973], takes evolutionary biology as its basis, with the selection, crossover and mutation of a population of individuals. Let $\mathcal{S}_i$ be an individual, that is, a sub-ensemble, and let $\mathcal{P} = \{\mathcal{S}_1, \ldots, \mathcal{S}_i, \ldots, \mathcal{S}_{N_{pop}}\}$ be a population of N_{pop} individuals. The first step of the genetic algorithm is the random generation of the first population (denoted $\mathcal{P}^0$). Each $\mathcal{S}_i$ randomly collects an arbitrary number of models of the ensemble $\mathcal{E}$. Then, three important steps generate the population $\mathcal{P}^{k+1}$ based on $\mathcal{P}^k$:

[32] 1. Selection: a few individuals are selected according to some method. In practice, we select half the best individuals with respect to the score.

[33] 2. Crossover: among the selected individuals, a crossover is carried out. Two parents $\mathcal{S}_a$ and $\mathcal{S}_b$ create two new children $\mathcal{S}_c$ and $\mathcal{S}_d$. All the models of $\mathcal{S}_a$ and $\mathcal{S}_b$ are randomly dispatched into $\mathcal{S}_c$ and $\mathcal{S}_d$. The list of models in an individual can be seen as its genetic print. A new population denoted $\tilde{\mathcal{P}}^{k+1}$ is generated with $N_{pop}/2$ parents and $N_{pop}/2$ children.

[34] 3. Mutation: each individual of the previous population $\tilde{\mathcal{P}}^{k+1}$ can mutate. In our case, a model can be replaced by another one, removed from an individual or added to an individual. These mutations constitute the new population $\mathcal{P}^{k+1}$.

[35] The operation is repeated until some stopping criterion has been satisfied, e.g., when a given number of iterations is reached. The final population contains many individuals that are better (with respect to the cost function) than those of the initial population. It is the best individual of the final population that is considered as the calibrated ensemble.

2.2.2.2. Simulated Annealing

[36] Simulated annealing, described by *Kirkpatrick et al.* [1983], is a basic optimization method inspired by a thermodynamic process. Each sub-ensemble of the search space is analogous to a state of some physical system.

[37] In our case, the first state is just a random generation of a sub-ensemble. The current state has a lot of neighbor states which correspond to the current state with a unit change, that is, a removed, added or replaced model in the sub-ensemble. Let $\mathcal{S}$ be the current sub-ensemble and $\mathcal{S}'$ be a neighbor sub-ensemble. $\mathcal{S}'$ is a new sub-ensemble which is randomly built from the current sub-ensemble with one removed, added or replaced model. In order to minimize (*resp.* maximize) a score J, two transitions to the neighbor are possible:

[38] 1. If the score $J(\mathcal{S}')$ is lower (resp. higher) than $\mathcal{S}$), then the current sub-ensemble moves to the neighbor sub-ensemble. $\mathcal{S}'$ becomes the current sub-ensemble and another neighbor is generated.

[39] 2. If the $J(\mathcal{S}')$ is greater (resp. lower) than $J(\mathcal{S})$, moving to $\mathcal{S}'$ is allowed to occur with an acceptance probability. This acceptance probability is equal to $\exp\left(-\frac{J(\mathcal{S}')-J(\mathcal{S})}{T}\right)$ (*resp.* $\exp\left(\frac{J(\mathcal{S}')-J(\mathcal{S})}{T}\right)$) where T is called temperature and is decreased after each iteration. A state movement is carried out if $u < \exp\left(-\frac{J(\mathcal{S}')J(\mathcal{S})}{T}\right)$ where u is a random number uniformly drawn from [0, 1]. At the beginning of the algorithm, the acceptance probability is high. Thus, the probability of switching to neighbor is higher than at the end of the algorithm.

[40] At the end of the process, the best state encountered in all the iterations, i.e., the best sub-ensemble, is taken as the calibrated ensemble.

3. Application to a 101-Member Ensemble

[41] We consider the 101-member ensemble, launched throughout the year 2001 over Europe and described in

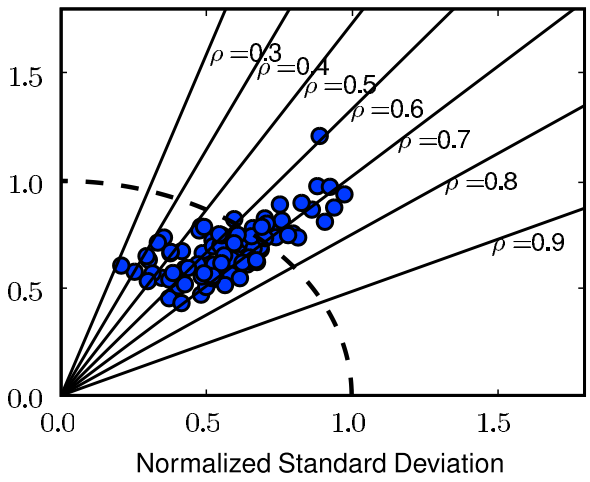

Figure 1. Taylor plots of ozone peak averaged over stations. The radial coordinate is the standard deviation normalized by the standard deviation of observations. The angles between the abscissa axis and the lines correspond to the arccosine of the correlation ρ between each simulation and observations.

detail by *Garaud and Mallet* [2010]. The ensemble was automatically generated for the simulation of ground-level ozone, with a horizontal resolution of half a degree. Each member of the ensemble is a unique combination of physical parameterizations, numerical schemes and input data. For instance, the members can differ in the chemical mechanism (RACM or RADM2), the computation of the vertical diffusion coefficient (Louis' or Troen&Mahrt's parameterizations), the vertical resolution (5 or 9 levels) or the perturbation of the meteorological fields (wind, temperature, etc.) and emission sources. About 30 alternatives are available for the generation of a member. The generated ensemble contains very different members and has a wide spread. The following subsections deal with the assessment of this ensemble and its calibration according to ensemble scores previously mentioned.

3.1. Evaluation of the Ensemble

[42] In this sub-section, we quickly review the performance of the models and then of the ensemble.

[43] The ensemble evaluation is carried out using the observation network Airbase (http://air-climate.eionet.europa.eu/databases/airbase/airbasexml/index_html). This database, managed by the European Environment Agency, provides ground-level ozone observations at 210 rural background, 702 rural, 647 suburban and 1324 urban stations across Europe.

[44] Stations that fail to provide observations at over 10% of all the dates considered are discarded as the scores at these stations may not reliable. In order to have stations which are representative of the ozone peak concentration at the model scale (half a degree in the horizontal), only rural and background stations are kept. There are about 123,000 observations for ozone peaks during the year 2001. Following usual recommendations [*Russell and Dennis*, 2000; *Hogrefe et al.*, 2001; *U.S. Environmental Protection Agency*, 1991], a cut-off is applied to the observations. Observations below 40 μg m^{-3} are discarded so as to focus on the most harmful concentrations.

3.1.1. Models Skills

[45] The different models show quite different skills and performances. The spatio-temporal mean of ground-level ozone peaks ranges from 60 to 130 μg m^{-3}. Their variability is also quite different because the global standard deviation of ozone peak simulations ranges between 17 and 44 μg m^{-3}.

[46] Figure 1 shows the performance, compared to the observations, of the 101 simulations in a single diagram. This Taylor diagram [*Taylor*, 2001] takes into account the standard deviation of the observations and the correlation between each simulation and the observations. The radial coordinate of the Taylor diagram corresponds to $\frac{\sigma_x}{\sigma_y}$ where σ_x is the empirical standard deviation $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\frac{1}{n}\sum_{j=1}^{n}x_j\right)^2}$ of the simulated sequence $(x_i)_{i=1,\ldots,n}$, and σ_y is the empirical standard deviation of the observed sequence $(y_i)_{i=1,\ldots,n}$. The azimuth is the arccosine of the correlation between $(x_i)_{i=1,\ldots n}$ and $(y_i)_{i=1,\ldots,n}$. The lower azimuth, the higher correlation between a simulation and the observations. A Taylor diagram shows the performance of an ensemble of simulations in term of correlation, the variability of each simulation compared with the observed variability, and the spread of these performances. Although a large number of simulations show less variability than the observations, a number of members still show good variability. The correlations range between 0.3 and 0.77.

[47] This shows that the ensemble has a strong variety and that the models can have very different statistical measures and performance. A few models have weak skill, i.e., a high RMSE (up to 29.6 μg m^{-3}) and a low correlation (down to 0.3). However these models should not be discarded because they can bring useful information. Figure 2 shows the number of times each model is closer to an observation than any other model. Most of the bars are close to the mean (1091 observations). Figure 2 shows that all the members give the closest concentrations to the observations for a

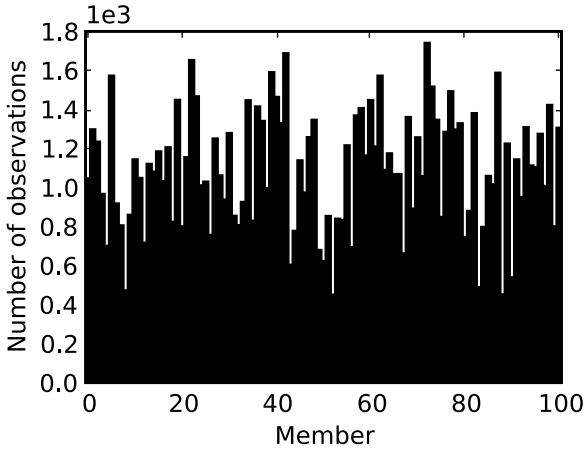

Figure 2. Best models count for ozone peaks on the network Airbase. A model is counted "best" when the discrepancy between the simulated concentration and the observation is minimal. The count is carried out for all observations.

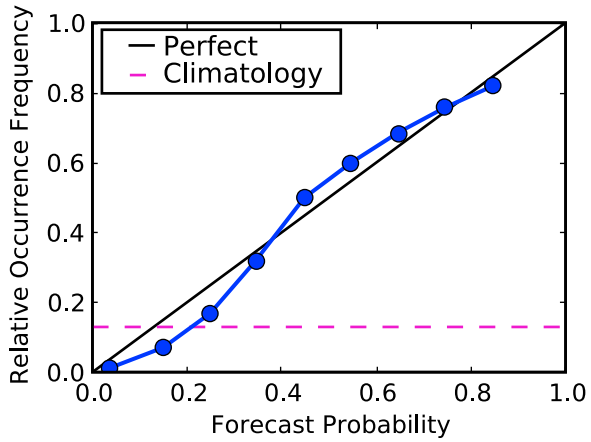

Figure 3. Reliability diagram of the ensemble for ozone peaks. The ozone concentration threshold is 120 μg m^{-3}. The black line corresponds to a perfect reliability diagram. The dashed horizontal line is the value of the climatological forecast.

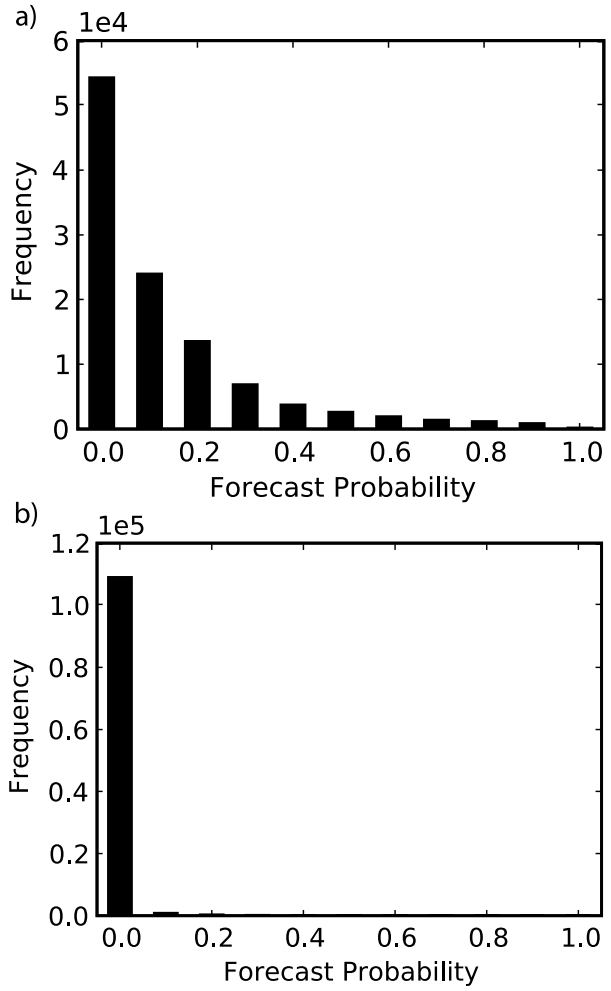

Figure 4. Sharpness histograms for two ozone concentration thresholds: (a) 120 μg m^{-3} and (b) 180 μg m^{-3}.

significant number of times. In the worst case, the count is about half the mean count. The worst model in terms of RMSE and correlation gives the closest concentrations to 1061 observations, which is about the average performance. This means that even if a member shows a bad performance on average, it still brings useful information in some regions and at some dates.

3.1.2. Ensemble Scores

3.1.2.1. Reliability Diagram

[48] Figure 3 shows the reliability diagram for the event $[O_3] \geq 120$ μg m^{-3}. The ensemble shows a reasonable performance since the diagram roughly follows the diagonal. Below the forecast probability 0.4, the ensemble overforecasts the event occurrence since the reliability curve is below the diagonal. On the other hand, the ensemble under forecasts the event occurrence when the forecast probabilities are greater than 0.4. The diagram shows that the ensemble has an acceptable resolution. An ensemble with lower resolution would have a flatter reliability diagram which would be close to the climatological forecast. Unfortunately, for an event based on a higher concentration, such as $[O_3] \geq 180$ μg m^{-3}, the ensemble leads to a poor reliability diagram. This can be explained by the very low occurrence of the event – about 0.6% of all cases – and by the sharpness histogram. Two sharpness histograms are shown in Figure 4 and represent the frequency of the forecast probabilities for the two previous events. The sharpness indicates the tendency of an ensemble to provide probabilities near 0 or 1. The forecast probabilities provided for the first event (120 μg m^{-3}) are quite frequent and close to 0. Thus, most of the time, no simulation exceeds the threshold, so that the ensemble gives a null probability of event occurring. For the threshold 180 μg m^{-3}, the sharpness histogram is even worse since over 98% of forecast probabilities are less than 0.1. As the number of forecast probabilities greater than 0.1 is so low, it seems difficult to correctly build a reliability diagram. Hence for the threshold 180 μg m^{-3}, the calibration cannot be carried out using the reliability diagram.

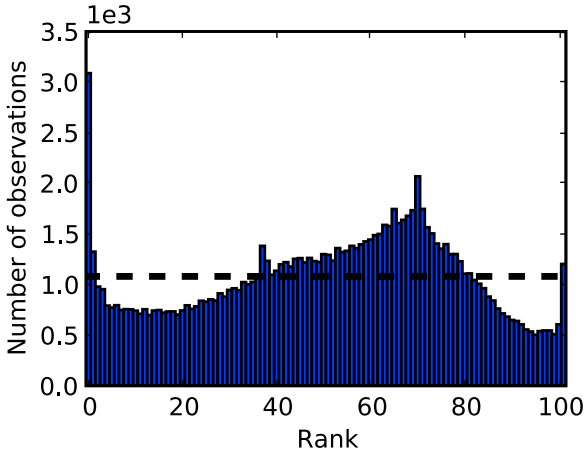

Figure 5. Rank histogram of the 101-member ensemble on network Airbase for ozone peaks. The horizontal dashed line corresponds to the ideal value for a flat rank histogram with respect to the number of members. The large number of observations on the left means there are many observations below the lower envelope of the ensemble.

Table 1. Brier Scores and Brier Skill Scores for the Event $[O_3] \geq 120\ \mu g\ m^{-3}$ for the Ensemble, the "Best" Model, With Respect to the Score, and the Climatological Forecast[a]

	Full Ensemble	Best Model	Climatology
Brier	$76\ 10^{-3}$	$95\ 10^{-3}$	$113\ 10^{-3}$
Brier skill	$32.7\ 10^{-2}$	$15.6\ 10^{-2}$	0.0
DRPS	$90.3\ 10^{-3}$	$124\ 10^{-3}$	$130\ 10^{-3}$

[a]The DRPS is computed with the threshold exceedances for 80, 100, 120, 140 and 160 $\mu g\ m^{-3}$.

3.1.2.2. Rank Histogram

[49] Figure 5 is the rank histogram of the 101-member ensemble for ozone peaks. The histogram does not show any extremely low or extremely high bar, but several bars have half the height they should have and several others are significantly higher than expected. The first bar, which corresponds to the number of observations below the lower envelope of ensemble, is especially high. It means that, at certain locations and dates, the spread of the ensemble is insufficient to cover the observations. The measure of the flatness described in the section 2.2.1.1 is 148.

3.1.2.3. Brier Score and DRPS

[50] The Brier score, Brier skill score and discrete ranked probability score are computed with the full ensemble, with the "best" model alone and with the climatological forecast. The "best" model will be the member from the full ensemble that minimizes or maximizes the given score. The climatological forecast is given by the all-year relative (observed) occurrence frequency of the event. These different scores are reported in Table 1. The DRPS is computed with the threshold exceedances for 80, 100,120, 140 and 160 $\mu g\ m^{-3}$.

[51] It is interesting to notice that the "best" model is always the same for all scores and corresponds to the model which has the smallest RMSE (20.5 $\mu g\ m^{-3}$). This "best" model is always better than the climatological forecast. It should, however, be noted that, first, one model can only provide probabilities equal to 0 or 1 and secondly, a large majority of the models have worse scores than the climatological forecast. For instance, over 77% of the models have a negative Brier skill score for the 120 $\mu g\ m^{-3}$ threshold exceedance. Whatever the score, the full ensemble always performs better than the "best" model. Consequently it seems that an ensemble is necessary to provide forecast probabilities which are more accurate than probabilities provided by a single model.

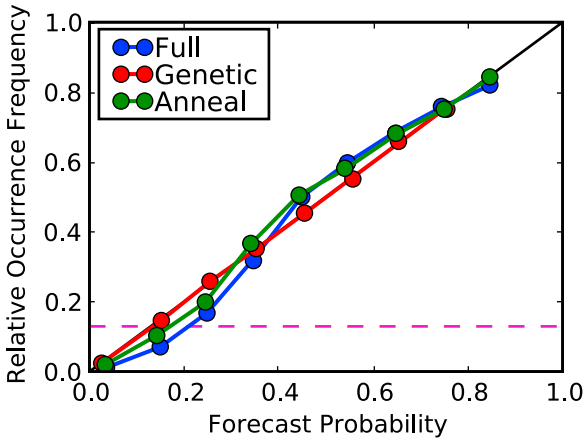

Figure 6. Calibrated reliability diagrams for the event $[O_3] \geq 120\ \mu g\ m^{-3}$ from the simulated annealing and the genetic algorithm. The dashed line corresponds to the value of the climatological forecast.

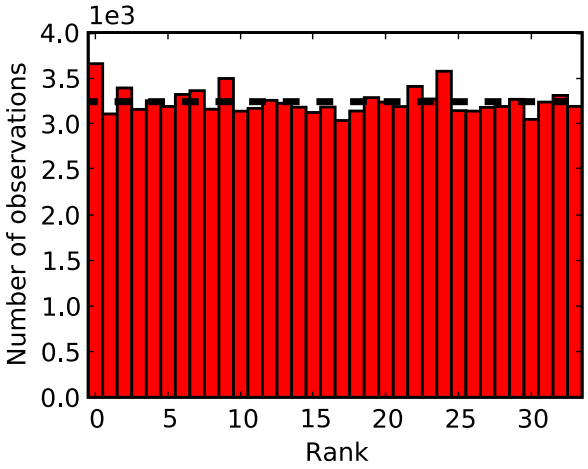

Figure 7. Rank histogram of the calibrated ensemble on network Airbase for ozone peaks. The horizontal dotted line corresponds to the ideal value for a flat rank histogram according to the number of members.

3.2. Calibration

3.2.1. Reliability Diagram

[52] We introduce the average probability $\overline{p}_k$ of all forecast probabilities lying in the interval $[p_{k-1}, p_k]$. As described in the section 2.2.1.2, a perfect reliability leads to $\overline{p}_k = O_k$. In order to have an optimized reliability diagram, the calibration method is therefore carried out with the mean squared error of the diagram. The score to minimize can be written as

$$\mathcal{C}_{rel} = \frac{1}{K}\sum_{k=1}^{K}(\overline{p}_k - O_k)^2. \tag{10}$$

[53] We consider the event $[O_3] \geq 120\ \mu g\ m^{-3}$, and we apply the genetic algorithm and the simulated annealing.

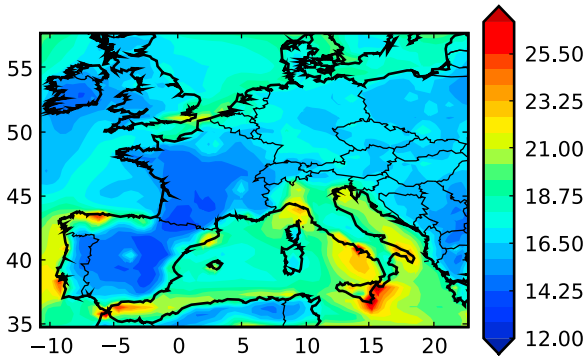

Figure 8. Monthly average of ozone uncertainty from a calibrated sub-ensemble for June 2001 across Europe ($\mu g\ m^{-3}$).

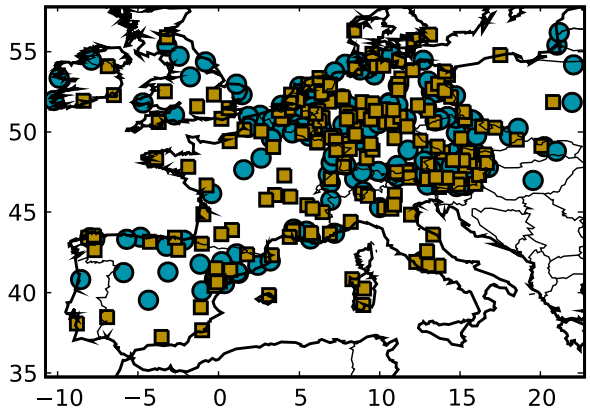

Figure 9. The two random sets of stations over Europe. These two sub-networks are used to assess the spatial robustness of the ensemble calibration method. The two sub-networks are a partition of the full network: each station of the full network belongs to one and only one sub-network.

Figure 6 shows the two resulting reliability diagrams. The calibrated diagrams are better than the reliability diagram of the full ensemble since they are closer to the diagonal. The 35-member calibrated ensemble from the genetic algorithm is very reliable and has a mean squared error lower than 10^{-5}. As the reliability is improved, the Brier skill score of the two calibrated sub-ensembles are equal to 34 10^{-2} and 35 10^{-2}, which represents slight improvements compared with the full ensemble. The Brier score decomposition shows that the reliability term is better after calibration whereas the resolution term is slightly worse. For the best calibrated sub-ensemble (genetic algorithm), the reliability term decreases by about 93% while the resolution term decreases by about 1%. *Candille and Talagrand* [2005] show that there is a compromise between reliability and resolution. Thus, resolution can be degraded when reliability is improved. Nevertheless, this calibration dedicated to improving reliability degrades resolution very slightly.

3.2.2. Rank Histogram

[54] We now apply the calibration with criterion (2) so as to get a flat rank histogram. Note that it is desirable to obtain a sub-ensemble with the largest number of models so that an accurate uncertainty estimation can be produced. It is possible to obtain a perfectly flat diagram with just one model, providing half the observations are below the model concentrations and half the observations are above; but one model cannot help in providing an uncertainty estimation.

[55] The calibration results depend on the height of the highest bar (here, the left bar) of the full-ensemble histogram. All observations with rank 0 (left bar) are below the lower envelope of the ensemble. For any sub-ensemble, the height of the left bar cannot be lower than the number r_0 of observations below the lower envelope. In a flat histogram, at best, the height of the left bar is still r_0 and all the bars have the same height. In this case, there cannot be more than 34 members (which is deduced from the total number of

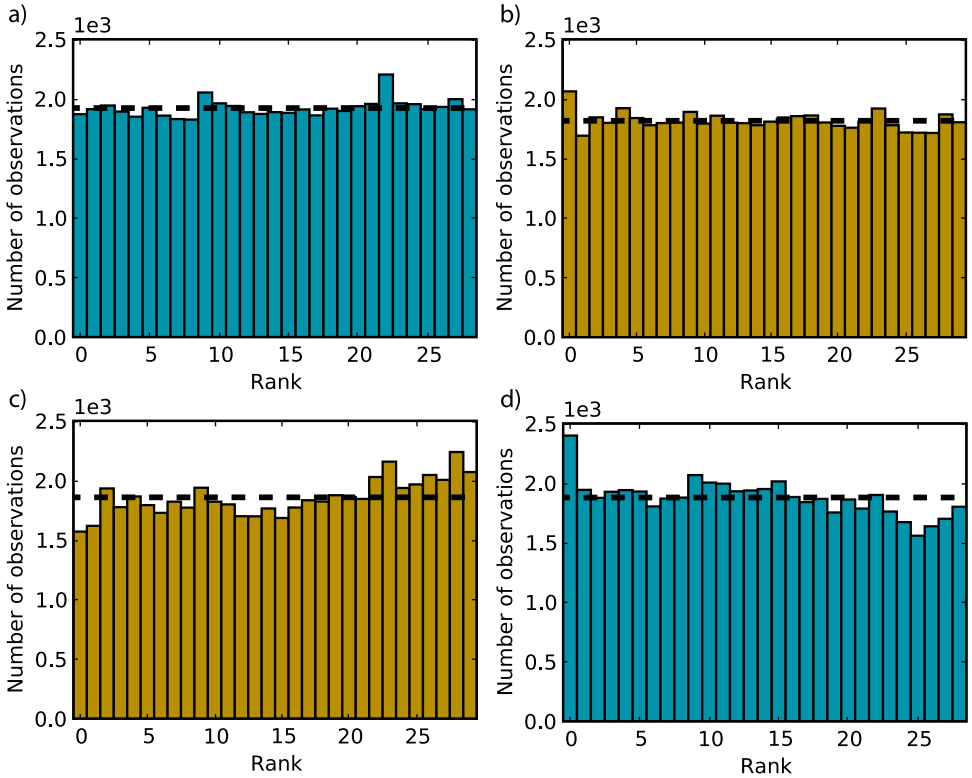

Figure 10. Rank histograms of the calibrated sub-ensembles on the two random sub-networks. The calibrated rank histograms of the (a) cyan and (b) yellow sub-ensembles. The rank histograms computed (c) from the yellow sub-ensemble on the cyan sub-network and (d) from the cyan sub-ensemble on the yellow sub-network.

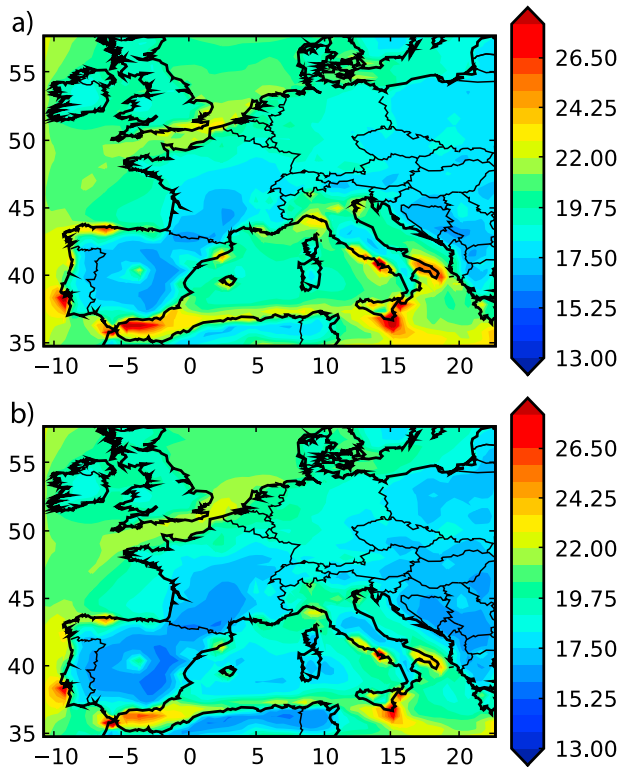

Figure 11. Temporal average of uncertainty estimation in μg m^{-3} from two sub-ensembles which were calibrated with two random sub-networks over Europe. Uncertainty map from (a) the cyan network and (b) the yellow network for June 2001.

observations divided by r_0). Figure 7 is the rank histogram of the calibrated sub-ensemble using simulating anneal. There are 33 members and the flatness score is about 6 instead of 148 for the full ensemble score.

[56] This calibrated sub-ensemble also improves the Brier scores and the DRPS. For the same events as before, the Brier skill score and DRPS respectively give 36 10^{-2} and 90 10^{-3}. It is interesting to notice that the reliability (from the DRPS decomposition (9)) is decreased by 90%, while the resolution remains unchanged. This is consistent with the fact that the rank histogram is an ensemble score which measures reliability.

3.2.3. DRPS

[57] The calibration according to the DRPS gives DRPS_{calib} = 66 10^{-3}. The DRPS of the full ensemble is reduced by 15%. The reliability part (see (9)) is reduced by 47% and the resolution part by 10%.

[58] For all ensemble scores, the calibration provides well balanced sub-ensembles. They always are better than the full ensemble, the best model or the climatology. The calibrated sub-ensembles also improve the reliability. However, the resolution essentially remains the same. As for the Brier score decomposition (8), the resolution term does not depend directly on the agreement between the forecast probability and the event occurrences. The improvement in the resolution depends on the definition of forecast probabilities bins described in paragraph 2.2.1.2 and [*Candille and Talagrand*, 2005]. The ensemble calibration essentially improves the quality of forecast probabilities, i.e., the reliability, rather than the variance of frequency occurrence O_k.

4. Uncertainty Estimation

[59] We now analyze the uncertainty estimation based on the sub-ensemble calibrated for the rank histogram. This calibration is chosen because it is related to the probability distribution of ozone concentrations, whereas the other scores are used to assess an ensemble for specific events.

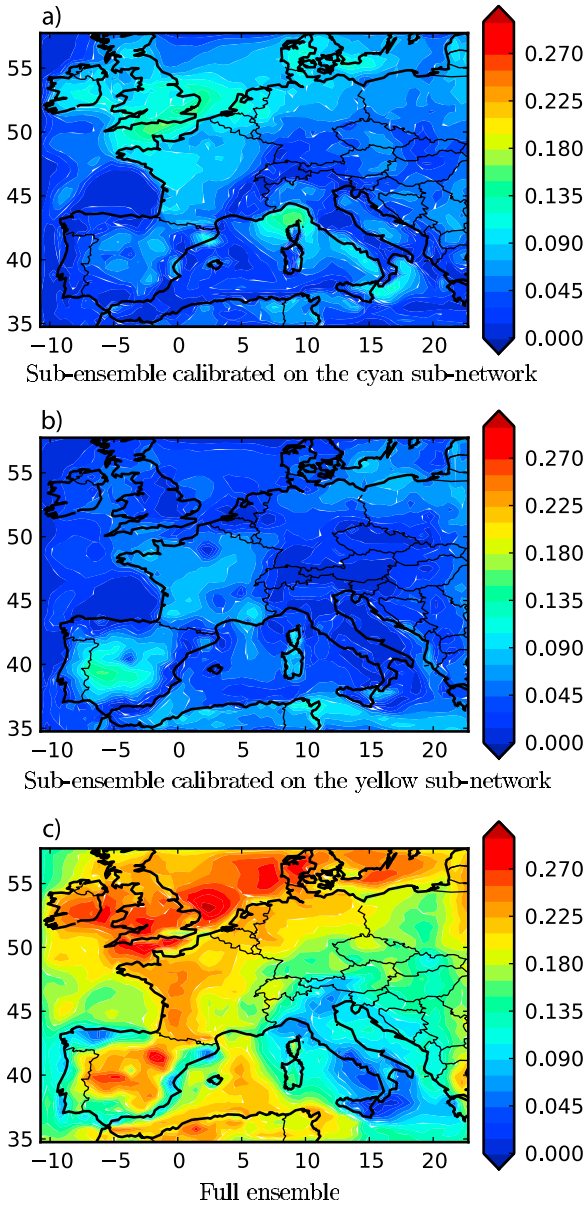

Figure 12. Relative discrepancy on uncertainty fields (averaged over June) between the sub-ensemble calibrated with all observations and (a) the sub-ensemble calibrated on the cyan sub-network, (b) the sub-ensemble calibrated on the yellow sub-network, and (c) the full ensemble. For example, the relative discrepancy (Figure 12c) is defined (pointwise) as the difference between the averaged uncertainty obtained with the full ensemble and the averaged uncertainty obtained with the calibrated sub-ensemble, divided by the latter.

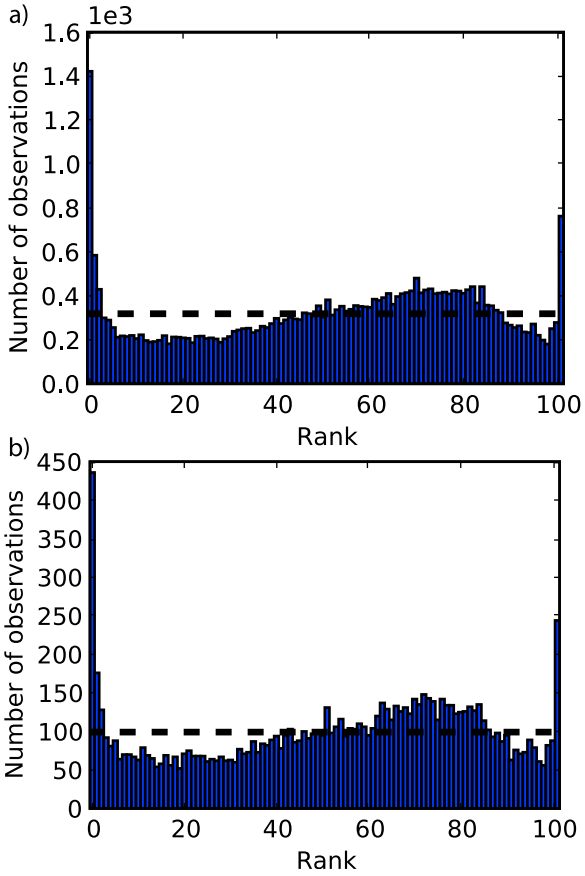

Figure 13. Rank histograms with a different number of observations: (a) about 32,500 and (b) 10,100 observations.

[60] The uncertainty can be estimated with the (empirical) standard deviation of the ensemble. A monthly average of the standard deviation of the calibrated ensemble is computed in each cell of the domain studied. Figure 8 shows the corresponding uncertainty map over Europe, averaged over June 2001. A higher ozone uncertainty appears along the south-coasts of Europe. This is consistent with a well-known difficulty of predicting ozone along the coasts, mainly because of poor representation of winds and turbulence in these areas.

[61] Before presenting further results, it is important to assess the robustness of the calibration method. One question is the spatial robustness. A calibrated sub-ensemble is spatially robust if it is still reliable at non-observed locations. In order to check this robustness, we randomly exclude stations from the calibration, and assess the calibration on the remaining stations.

[62] Figure 9 shows all observation stations previously used to compute the ensemble scores and to calibrate the ensemble. This network is randomly split into two sub-networks (cyan and yellow). The rank histogram calibration is then carried out on each sub-network, that is, using only the observations of the sub-network. Figure 10 shows four rank histograms for the two calibrated sub-ensembles. At the top of the figure, the calibrated rank histograms are shown, each computed with the observations used for their calibration. At the bottom, the rank histograms are computed using the observations of the other sub-network. The rank histograms are almost flat, which shows that the calibration is robust. It is noteworthy that the two sub-ensembles have a similar number of members (27 and 28 members for the "cyan" and "yellow" sub-ensembles, respectively).

[63] We can now compare the uncertainty estimation maps from the two previous calibrated sub-ensembles. Figure 11 shows the uncertainty estimation of the two calibrated sub-ensembles from the two previous random sub-networks. The spatial structures are similar. The high and low uncertainty values are located at the same places. In Figure 12, these uncertainty maps are also compared with the uncertainty map obtained after calibration with all observations. The relative difference between these maps is about 3% on average, and marginally exceeds 10%. For reference, the figure also shows the relative difference with the uncertainty derived from the full ensemble.

[64] Besides spatial robustness, the previous results also show that here, half observations are sufficient to calibrate an ensemble and estimate uncertainties. This raises the question of how many observations are needed for the calibration. An experiment was carried out to estimate this number. First, a rank histogram is computed for the full ensemble with about 30,000 hourly observations. These observations are selected arbitrarily. Then, observations are randomly removed and the rank histogram is computed again. After a few iterations, we can compare several rank histograms with a different number of observations. Figure 13 shows two rank histograms of the full ensemble with about 32,500 observations and about 10,100 observations. Their

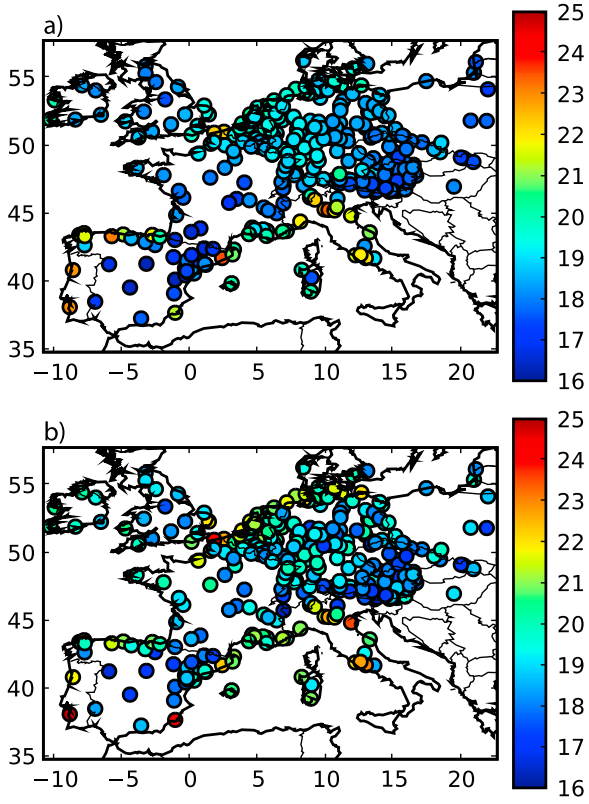

Figure 14. Uncertainty estimations at stations (a) for the reference calibration and (b) for the calibration with perturbed simulations ($\varepsilon = 0.13$).

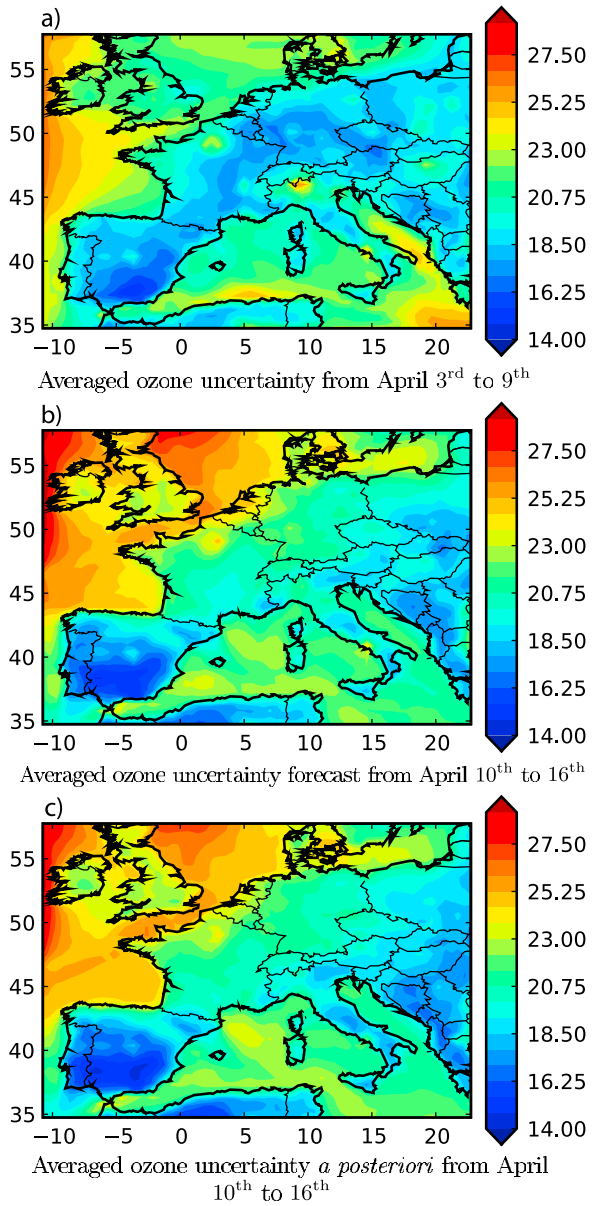

Figure 15. Comparison of ozone uncertainty maps averaged over one week in μg m^{-3}. (a) The uncertainty estimation during the learning period (from 3rd to 9th April 2001), (b) the uncertainty forecast (10th to 16th April 2001), and (c) the a posteriori uncertainty.

shapes are very similar. Below 8000 observations, the shape of the rank histogram starts changing. So we conclude that 8000–10,000 observations are required to assess the quality of the 101-member ensemble.

[65] A similar experiment was carried out to determine the number of observations needed for the calibration to be reliable. The full ensemble is calibrated a few times with a total number of observations (initially 32,500) divided by 2, 3, 5, 8 and 13. The calibrated ensembles contain a similar number of members, ranging from 22 to 27. The rank histograms for the calibrated ensembles are then computed, each time with the observations used in the calibration. The rank histograms remain flat in every case. The uncertainty estimations starts depending on the number of observations when there are fewer than 8000–10,000 observations.

[66] Another question is the impact of observational errors on the calibration and on the uncertainty estimation [*Anderson*, 1996; *Hamill*, 2001]. The rank histogram checks whether two random variables sample the same distribution. Noise in the observations should therefore be added to ensemble so that we can check the ensemble samples the real uncertainty without observation noise. Let $x_i^m(t)$ be the simulated concentration at station i and date t for the model m. We assume that observational errors do not depend on the station and date. We introduce the perturbed concentrations $\hat{x}_i^m(t) = x_i^m(t)(1 + \alpha_i^m)$ where α_i^m follows a uniform distribution on the interval $[-\varepsilon, \varepsilon]$. This form allows us to introduce a noise relative to the concentration, which is a usual feature for ozone observations. Based on work by *Airparif* [2007], $\varepsilon \simeq 0.13$ for ozone peak concentrations measured over the year 2009 at about 30 stations from the Airparif monitoring network (in the Paris region). This noise is introduced before the calibration. The calibrated ensemble

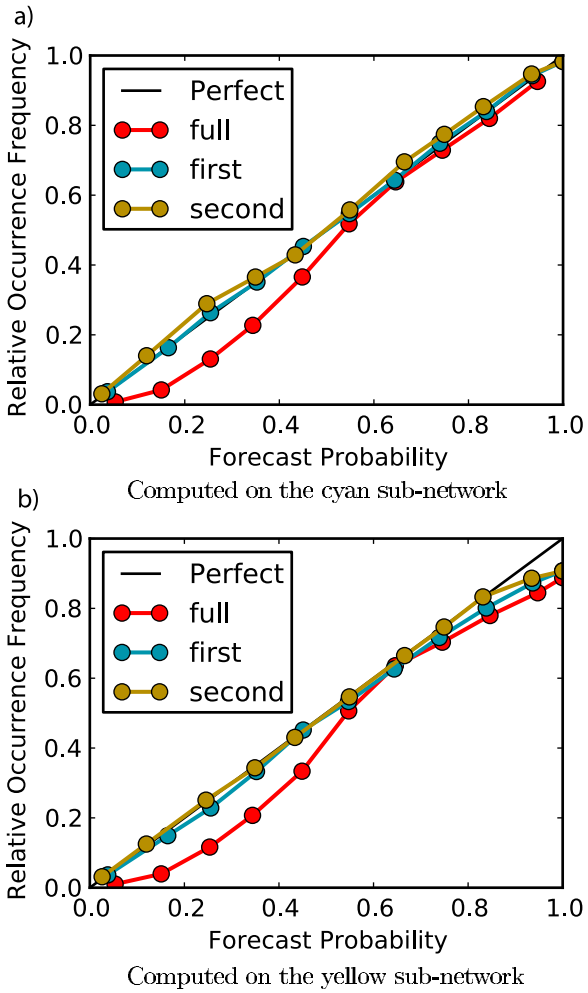

Figure 16. Reliability diagrams for $[O_3] \geq 100$ μg m^{-3} of the calibrated sub-ensembles and the full ensemble. (a) The reliability diagrams are computed on the cyan sub-network. (b) The reliability diagrams are computed on the yellow sub-network.

Table 2. Brier Skill Scores of the Full Ensemble and the Calibrated Sub-ensembles[a]

Threshold Exceedance	$[O_3] \geq$ 80 μg m^{-3}	$[O_3] \geq$ 100 μg m^{-3}	$[O_3] \geq$ 120 μg m^{-3}
Full ensemble	0.35	0.37	0.34
Cyan calibrated sub-ensemble	0.40	0.46	0.44
Yellow calibrated sub-ensemble	0.40	0.46	0.44

[a]The scores are computed using the observations of cyan sub-network.

with perturbation ($\varepsilon = 0.13$) shows a flat rank histogram, and the resulting uncertainty estimations are plotted in Figure 14. The values and spatial patterns of the standard deviation are very similar to those of the calibration without perturbations. The observation errors therefore seem to have a limited impact on the calibration.

[67] Finally, we investigate the robustness of the calibration over time. A calibration is carried out during a learning period, and the relevance of this calibration is evaluated for a forecast period. The sub-ensemble selected based on the learning period is referred to as an a priori sub-ensemble. The quality of the forecast is measured by comparing the a priori sub-ensemble and the a posteriori sub-ensemble that is calibrated over the forecast period.

[68] The learning period is a week, from April 3rd to April 9th, with 50,000 hourly ozone observations. It is an arbitrary chosen period. The forecast period ranges from April 10th to April 16th. Figure 15 shows the uncertainty map computed during the learning period and the forecast uncertainty map. These maps clearly show different patterns, e.g., with higher forecast uncertainties over the North Sea, over France and Germany, and with lower forecast uncertainties over several parts of the Mediterranean Sea. This, and tests not reported here, show that the uncertainty estimations can vary strongly over time. Figure 15 also shows the a posteriori uncertainty map. The forecast and a posteriori maps essentially show the same patterns and uncertainty levels. This means that, despite the significant variation in time, the calibration seems robust over time. Here the calibration can be used to forecast the uncertainties for a few days. The root mean square error between the forecast and a posteriori maps (daily averages), divided by the mean of the a posteriori map, is equal to about 5% over each of the next six days. It is noteworthy that the learning period should be long enough—two-day or four-day periods do not appear to be long enough to ensure a good forecast.

5. Risk Assessment and Probabilistic Forecast

[69] In order to check that the calibration can help in risk assessment and in forecasting a given event, the same tests as in the previous section are carried out with the Brier skill score and the reliability diagram instead of the rank histogram.

[70] Figure 16 shows, for each sub-network, reliability diagrams for calibrated sub-ensembles and the full ensemble. Any sub-ensemble calibrated on one sub-network performs well on the other sub-network.

[71] The same conclusion can be drawn from the Brier skill score calibration. Table 2 shows the Brier skill scores of the full ensemble and calibrated sub-ensembles computed on the cyan sub-network for three different thresholds — 80, 100 and 120 μg m^{-3}. Whatever the threshold exceedance, the calibrated sub-ensembles perform significantly better than the full ensemble. The sub-network over which the calibration was carried out does not impact the results.

[72] According to these results, the calibrations based on the reliability diagram and the Brier skill score seem spatially robust.

[73] In order to assess the temporal robustness, we select arbitrarily the learning period from May 31th to June 6th and rely on the corresponding calibrated sub-ensemble to forecast the period from June 7th to June 13th. Figure 17 shows the reliability diagrams of the full ensemble, the a priori calibrated sub-ensemble and the a posteriori calibrated sub-ensemble, for the threshold exceedance $[O_3] \geq 100$ μg m^{-3}. The a priori sub-ensemble performs better than the full ensemble, but its reliability diagram is deteriorated compared to the a posteriori sub-ensemble. Note that the forecast period is long (7 days) because the reliability diagram requires a significant amount of data to be computed. It is possible that the results would be better if the diagram could be computed with the observations of the very first forecast days only.

[74] The Brier skill scores in the same forecast period are 0.18, 0.27 and 0.25 for the full ensemble, the a posteriori sub-ensemble and the a priori sub-ensemble, respectively. It shows that the calibration can be relevant in the context of probabilistic forecast.

6. Conclusion

[75] The work presented in this paper relies on a 101-member ensemble that was automatically generated on the Polyphemus platform. This large ensemble is evaluated for uncertainty estimation and for probabilistic forecasts. The tests show that about 10,000 observations are required to properly evaluate the 101-member ensemble. A calibration method is designed to select a sub-ensemble from the full ensemble that better estimates the uncertainties.

[76] Several calibrations for different ensemble scores are carried out and show significant improvements in the ensemble scores. An almost perfect reliability diagram and a

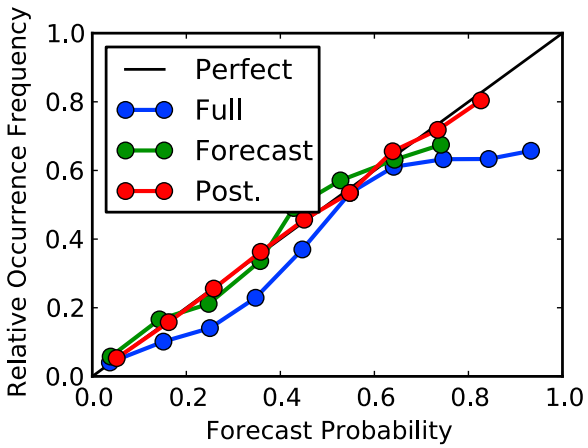

Figure 17. Reliability diagrams for $[O_3] \geq 100$ μg m^{-3} of the full ensemble (cyan), the a posteriori calibrated sub-ensemble (red) and the a priori calibrated sub-ensemble (green). This is based on observations from June 7th to June 13th.

very flat rank histogram can result from the calibration. We note that observation errors have a slight impact on calibration, since uncertainty maps with and without observation errors have the same pattern. The quality of the spatial distribution of the uncertainty estimation is assessed by a cross validation. Again, the calibration seems robust as the uncertainty maps are reasonably sensitive to the observation network. Finally, we show that the method can be applied in a forecasting context. The calibration can be carried out on a learning period, and the resulting sub-ensemble is able to estimate the uncertainties in the subsequent period almost as well as the sub-ensemble calibrated on this subsequent period.

[77] It would therefore be a natural next step to apply the method proposed here in operational conditions, including for aerosols for which the number of available observations may be significantly lower. A question is how much the proposed approach can help forecast threshold exceedances. The results show that the scores associated with such forecasts are improved, but the impact in an operational platform for decision making has yet to be assessed.

[78] The complexity of the method mainly lies in the automatic generation of a large ensemble in which many sources of uncertainties are taken into account. An open question is what ensemble design should be considered for uncertainty estimation and probabilistic forecasting. This question is especially important when considering forecasts because the sub-ensemble selected over one period should still represent the right uncertainty sources in another period. Monte Carlo simulations, for instance, are easier to carry out, but they might miss important uncertainty sources coming from the model formulation itself.

[79] Further work should address the partition of the uncertainty sources in order to better identify modeling errors, representativeness errors and measurement errors. Also the spatial and temporal correlations in the errors should be evaluated.

[80] **Acknowledgments.** We would like to thank to Hélène Marfaing and Christophe Debert from Airparif for their very useful studies and their data about measurement uncertainties. We thank Richard James for proofreading the paper.

References

Airparif (2007), Guide pratique d'utilisation pour l'estimation de l'incertitude de mesure des concentrations en polluants dans l'air ambiant, *Tech. Rep. V. 9*, Paris.

Anderson, J. L. (1996), A method for producing and evaluating probabilistic forecasts from ensemble model integrations, *J. Clim.*, *9*(7), 1518–1530.

Beekmann, M., and C. Derognat (2003), Monte Carlo uncertainty analysis of a regional-scale transport chemistry model constrained by measurements from the atmospheric pollution over the Paris area (ESQUIF) campaign, *J. Geophys. Res.*, *108*(D17), 8559, doi:10.1029/2003JD003391.

Brier, G. W. (1950), Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.*, *78*(1), 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Candille, G., and O. Talagrand (2005), Evaluation of probabilistic prediction systems for a scalar variable, *Q. J. R. Meteorol. Soc.*, *131*, 2131–2150, doi:10.1256/qj.04.71.

Crosby, J. L. (1973), *Computer Simulation in Genetics*, John Wiley, Hoboken, N. J.

Delle Monache, L., and R. B. Stull (2003), An ensemble air-quality forecast over western Europe during an ozone episode, *Atmos. Environ.*, *37*, 3469–3474.

Delle Monache, L., X. Deng, Y. Zhou, and R. B. Stull (2006a), Ozone ensemble forecasts: 1. A new ensemble design, *J. Geophys. Res.*, *111*, D05307, doi:10.1029/2005JD006310.

Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull (2006b), Probabilistic aspects of meteorological and ozone regional ensemble forecasts, *J. Geophys. Res.*, *111*, D24307, doi:10.1029/2005JD006917.

Epstein, E. S. (1969), A scoring system for probability forecasts of ranked categories, *J. Appl. Meteorol. Climatol.*, *8*(6), 985–987.

Fraser, A., and D. Burnell (1970), *Computer Models in Genetics*, McGraw-Hill, New York.

Garaud, D., and V. Mallet (2010), Automatic generation of large ensembles for air quality forecasting using the Polyphemus system, *Geosci. Model Dev.*, *3*(1), 69–85.

Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, *129*(3), 550–560.

Hamill, T. M., and S. J. Colucci (1997), Verification of Eta/RSM short-range ensemble forecasts, *Mon. Weather Rev.*, *125*, 1312–1327, doi:10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

Hanna, S. R., J. C. Chang, and M. E. Fernau (1998), Monte Carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables, *Atmos. Environ.*, *32*(21), 3619–3628.

Hanna, S. R., Z. Lu, H. C. Frey, N. Wheeler, J. Vukovich, S. Arunachalam, M. Fernau, and D. A. Hansen (2001), Uncertainties in predicted ozone concentrations due to input uncertainties for the UAM-V photochemical grid model applied to the July 1995 OTAG domain, *Atmos. Environ.*, *35*(5), 891–903.

Hogrefe, C., S. T. Rao, P. Kasibhatla, W. Hao, G. Sistla, R. Mathur, and J. McHenry (2001), Evaluating the performance of regional-scale photochemical modeling systems: Part II—Ozone predictions, *Atmos. Environ.*, *35*, 4159–4174.

Hopson, T. M., and P. J. Webster (2010), A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07, *J. Hydrometeorol.*, *11*(3), 618–641, doi:10.1175/2009JHM1006.1.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983), Optimization by simulating annealing, *Science*, *220*(4598), 671–680.

Mallet, V., and B. Sportisse (2006), Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: An ensemble approach applied to ozone modeling, *J. Geophys. Res.*, *111*, D01302, doi:10.1029/2005JD006149.

McKeen, S., et al. (2005), Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004, *J. Geophys. Res.*, *110*, D21307, doi:10.1029/2005JD005858.

Murphy, A. H. (1971), A note on the ranked probability score, *J. Appl. Meteorol. Climatol.*, *10*(2), 155–156.

Murphy, A. H. (1973), A new vector partition of the probability score, *J. Appl. Meteorol. Climatol.*, *12*(4), 595–600.

Pinder, R. W., R. C. Gilliam, K. W. Appel, S. L. Napelenok, K. M. Foley, and A. B. Gilliland (2009), Efficient probabilistic estimates of surface ozone concentration using an ensemble of model configurations and direct sensitivity calculations, *Environ. Sci. Technol.*, *43*(7), 2388–2393.

Russell, A., and R. Dennis (2000), NARSTO critical review of photochemical models and modeling, *Atmos. Environ.*, *34*, 2283–2234.

Talagrand, O., R. Vautard, and B. Strauss (1999), Evaluation of probabilistic prediction system, paper presented at Workshop on Predictability, ECMWF, Reading, U. K.

Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res.*, *106*(D7), 7183–7192, doi:10.1029/2000JD900719.

U.S. Environmental Protection Agency (1991), Guideline for regulatory application of the urban airshed model, *Tech. Rep. EPA-450/4-91-013*, Research Triangle Park, N. C.

van Loon, M., et al. (2007), Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble, *Atmos. Environ.*, *41*, 2083–2097.

Vautard, R., et al. (2009), Skill and uncertainty of a regional air quality model ensemble, *Atmos. Environ.*, *43*, 4822–4832, doi:10.1016/j.atmosenv.2008.09.083.

Wilks, D. S. (2005), *Statistical Methods in the Atmospheric Sciences*, *Int. Geophys. Ser.*, vol. 100, 2nd ed., Academic, San Diego, Calif.

D. Garaud, CEREA, Université Paris-Est, 6–8 Av. Blaise Pascal, Cité Descartes, Champs-Sur-Marne, F-77455 Marne La Vallée CEDEX 2, France. (damien.garaud@cerea.enpc.fr)

V. Mallet, Paris-Rocquencourt Research Center, INRIA, Domaine de Voluceau, Rocquencourt, BP 105, F-78153 Le Chesnay, France. (vivien.mallet@inria.fr)